STATISTICAL RESEARCH REPORT
Institute of Mathematics
University of Oslo

No 5
September 1974

SIGNIFICANCE TESTING IN MULTIPLE STATISTICAL INFERENCE

by

Erling Sverdrup

Significance testing in multiple statistical inference.

by

Erling Sverdrup

1. Introduction. The meaning of significance testing has occasionally been the concern of statisticians, some of whom have even found it contradictory to test nullhypotheses which are known to be wrong, and which would have been rejected anyhow if the material was large enough. It will be the contention in this paper that in many situations the meaning of a significance testing is only properly understood in the context of a multiple inference problem and that only in such a context can the properties of the test be judged.

It seems to be specially in connection with chi-square-goodness-of-fit tests in multinomial situations that the problem has attracted attention. However, the problem is really quite general, and in this paper linear normal situations will also be considered.

No new methods of statistical inference and no new properties of known methods will be developped. The main purpose of this paper will be to clarify matters which some statisticians have been concerned about and to point out the relevance of some of the results obtained in statistical inference theory in the last decades. The present author feels that this circumstance, as well as the fact that the mathematics is elementary, occasions no excuses on his part.

2. Testing of one parameter.

a. Student's hypotheses: Suppose that $X_1, X_2, \ldots, X_n$ are independent normal (ξ,σ). We want to decide whether $\xi \leqq$ or $\geqq 0$, i. e. we have a choice between three decisions: "$\xi \leqq 0$", "$\xi \geqq 0$", "no inference" . The choice of inference method amounts to choice of acceptance regions B_1, B_2, B_3, respectively, for the three decisions. These regions are subsets of the sample space of all $x = (x_1,\ldots,x_n)$.

We could require of the method (B_1, B_2, B_3) .

(i) The level should be ε , i.e. the probability of wrong decision should be at most ε $(< \frac{1}{2})$,

$$\begin{aligned} &Pr(B_1|\xi,\sigma) \leq \varepsilon \quad \text{for} \quad \xi > 0 \\ &Pr(B_2|\xi,\sigma) \leq \varepsilon \quad \text{for} \quad \xi < 0 \end{aligned} \tag{1}$$

(ii) It should be performance unbiased, i. e.

$$\Pr(B_1|\xi,\sigma) \geq \varepsilon \quad \text{for} \quad \xi < 0$$
$$\Pr(B_2|\xi,\sigma) \geq \varepsilon \quad \text{for} \quad \xi > 0 \qquad (2)$$

(iii) Among all methods satisfying (i) and (ii) it should maximize $\Pr(B_1|\xi,\sigma)$ for all (ξ,σ) with $\xi < 0$ and maximize $\Pr(B_2|\xi,\sigma)$ for all (ξ,σ) with $\xi > 0$.

The almost unique method satisfying these requirements is the Student test,

$$B_1 = (\bar{X} < -\frac{tS}{\sqrt{n}}),\ B_2 = (\bar{X} > \frac{tS}{\sqrt{n}}),\ B_3 = (-\frac{tS}{\sqrt{n}} \leq \bar{X} \leq \frac{tS}{\sqrt{n}}) , \qquad (3)$$

where /t is the $(1-\varepsilon)$ - fractile of the Student distribution with n-1 degrees of freedom, $\bar{X}$ is the sample mean and S^2 is the usual unbiased estimate of σ^2 . This result follows immediately from well known optinum properties of the appropriate Student test for the two hypotheses $\xi \leq 0$ and $\xi \geq$ respectively.

Nothing can, of course, prevent us from performing the test in the following manner. Ascertain first if $\bar{X}$ is "significantly different from 0", i. e. if $|\bar{X}| > \frac{tS}{\sqrt{n}}$, - that is what is called "testing the null-hypothesis" - , assert then that $\xi \leq 0$ or $\xi \geq 0$ according as $\bar{X} < 0$, or $\bar{X} > 0$. There is of course nothing illogical in such a procedure even if the precise value $\xi = 0$ is excluded or unlikely a priori. On the contrary, the statistician ought to perform such a significance test if he wants to controle the chances of making errors and the sensitivity of the test as described above in items (i) - (iii). It is also clear that the method has tolerably good performance if (admittedly rather artificially) it should be a priori known that $\xi \notin (-\Delta,\Delta)$, $\Delta > 0$. "Testing $\xi = 0$" is justified under all circumstances, the purpose being to lay a possible foundation for asserting that $\xi \leq$ or ≥ 0 according as $\bar{X} <$ or > 0 . That similar interpretations of the purpose of many significance testings can be given, will be demonstrated below.

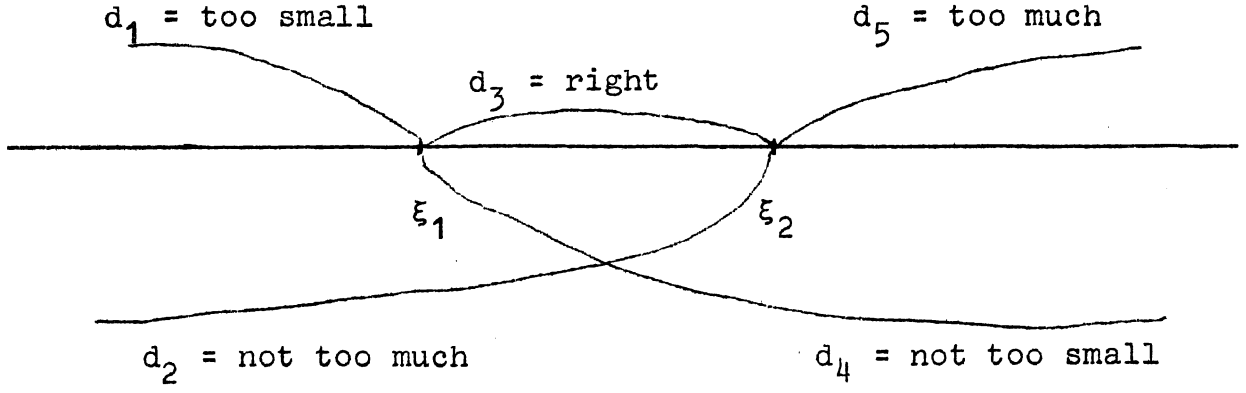
Of course, the power of the test, i. e. the probability of significance, is

$$\beta(\frac{\xi}{\sigma}) = G_{n-1}(-t,\xi\sqrt{n}/\sigma) + 1-G_{n-1}(t,\sqrt{n}/\sigma) \ , \qquad (4)$$

where $G_{n-1}(t,\lambda)$ is the cumulative Student distribution with $n-1$ degrees of freedom and eccentricity λ. This quantity tends to 1 when $n \to \infty$ if $\xi \neq 0$. Thus if n is large we would almost certainly discover the truth about ξ, whether ξ is $<$ or > 0. This state of affairs should only make the statisticians happy.

One might question if it is appropriate to keep the level ε constant as n increases, which would mean that prosperity on observations would unilaterally be beneficial to the sensitivity of the test. It is natural to argue that a large number of observations should mean reliable inferences, i. e. a low value of ε. To find the right balance between reliability $1 - \varepsilon$ and power β has been discussed by many authors, among others Erich Lehmann [1] and Anders Hald [2] who have approached the problem in different manners.

Some statisticians, who have found it meaningless to test the nullhypothesis, have defined a neighbourhood around the nullhypothesis, which in the present example would have the form $\xi_1 < 0 < \xi_2$. The idea would then presumably be to make a choice between decisions "$\xi_1 < \xi < \xi_2$", "$\xi < \xi_1$", "$\xi > \xi_2$". Now, if the possibility of not saying anything should be open to the statistician, then it should also be allowed to make the decisions "$\xi < \xi_2$" and "$\xi > \xi_1$", thus making it possible to make more or less daring decisions. We have a multiple location problem with decisions d_1 = "$\xi < \xi_1$" = "too small", d_2 = "$\xi < \xi_2$" = "not too much", d_3 = "$\xi_1 < \xi < \xi_2$" = "just about right", d_4 = "$\xi > \xi_1$" = = "not too small", d_5 = "$\xi > \xi_1$" = "too much", d_6 = "$-\infty < \xi < \infty$" = = "no statement". (The interpretations "bad","not good", "neither good nor bad", "not bad", "good" may also be appropriate).



The choice of procedure consists in partioning the sample space (of all X) in acceptance regions $B_1 \ldots, B_6$ for the six decisions $d_1, \ldots, d_6$. One such procedure is described in the table below.

Multiple location method.

			Loss $L_i(\xi)$				
i	Decisions d_i	Acceptance regions B_i	$\xi<\xi_1$	$\xi=\xi_1$	$\xi_1<\xi<\xi_2$	$\xi=\xi_2$	$\xi>\xi_2$
1	Too small	$\bar{X} \leq \xi_1 - V$	0	a	a+b	2a+b	2a+2b
2	Not too much	$\xi_1 - V < \bar{X} \leq \min(\xi_1 + V, \xi_2 - V)$	b	0	b	a+b	a+2b
3	Right	$\xi_1 + V \leq \bar{X} \leq \xi_2 - V$	a+b	a	0	a	a+b
4	Not too small	$\max(\xi_1 + V, \xi_2 - V) < \bar{X} < \xi_2 + V$	a+2b	a+b	b	0	b
5	Too much	$\bar{X} \geq \xi_2 + V$	2a+2b	2a+b	a+b	a	0
6	No statement	$\xi_2 - V < \bar{X} < \xi_1 + V$	2b	b	2b	b	2b

Here $V = tS/\sqrt{n}$, where S and t are as defined above. The method has the property that the probability of a wrong decision is at most 2ε. Note that if $n \geq 4t^2S^2/(\xi_2 - \xi_1)^2$ then the bold decision $\xi_1 < \xi < \xi_2$ is possible whereas the cautious "no statement" is excluded. If the inequality is reversed, the opposite is the case. If σ^2 had been known, and hence S^2 replaced by σ^2 in V, then one of these two decions, d_3 or d_6, could have been excluded a priori.

The merits of a method $\delta = (B_1, \ldots, B_6)$ could also be judged from the performance function

$$\beta_i(\xi, \sigma; \delta) = \Pr(\text{accepting } d_i);\ i = 1, \ldots, 6 \qquad (5)$$

A loss function $L_i(\xi, \sigma) \geq 0$ may be introduced, representing the loss inflicted by making decision d_i when (ξ, σ) is the true parameter value. The "risk" by using δ is then the expected value of the loss

$$r(\xi, \sigma; \delta) = \sum_{i=1}^{6} L_i(\xi, \sigma)\beta_i(\xi, \sigma; \delta) \qquad (6)$$

Since $L_i(\xi,\sigma)$ may be thought of as the "distance" between the parameter value (ξ,σ) and the decision d_i, then

$$r(\xi',\sigma';\xi,\sigma;\delta) = \sum_{i=1}^{6} L_i(\xi;\sigma')\beta_i(\xi,\sigma;\delta) \qquad (7)$$

represents the expected distance of the decision to the wrong parameter value $(\xi;\sigma')$, when (ξ,σ) is the true value. Since this quantity (the "anti-risk") should be large, whereas the risk (6) should be small, it would perhaps not be unreasonable to require of δ that

$$r(\xi',\sigma';\xi,\sigma;\delta) \geq r(\xi,\sigma;\delta) \qquad (8)$$

for all ξ',σ',ξ,σ. Such a δ is said to be "risk unbiased". It follows from results due to Lehmann [3] that if the loss function is given by $L_i(\xi,\sigma) = L_i(\xi)$ in the table above with $a+b = b\varepsilon$ and $a > b$, then the procedure δ_0, given in the table, uniformly minimizes the risk among all risk unbiased procedures δ. (The loss admittedly has some anomalous features, but seems on the whole to be reasonable).

b. Independence in double dichotomy.

We are interested in making inferences about the dependence of two factors A and B. The following frequency table is at our disposal

	B_1	B_2	
A_1	X	M-X	M
A_2	L-X	n-L-M+X	n-M
	L	n-L	n

(9)

Thus among the total number of observation n, there are X observations with $A_1 \cap B_1$, M with A_1 and L with B_1. We assume that the n observations form a multinomial sequence of trials with $p_{ij} = P_r(A_i \cap B_j)$; $i,j = 1,2$; in each trial $(\sum_{i,j} p_{ij} = 1)$.

The factors are independent if $\xi = p_{11} - (p_{11} + p_{12})(p_{11} + p_{21}) = 0$, or equivalently $\lambda = p_{11}p_{12}/(p_{12}p_{21}) = 1$. The dependence is negative or positive according as $\xi <$ or > 0; or equivalently according as $\lambda <$ or > 1.

We want to make one of the three decisions "$\xi < 0$" , "$\xi > 0$" , "no statement" ; and we require of our acceptance regions B_1, B_2, B_3 that they shall have analogous optimum properties to those specified in the Student case, i. e. (i), (ii) and (iii) above with ξ having the new meaning and σ replaced by $p = (p_{11},\ldots, p_{22})$. The optimum solution is roughly the same as the Irwin - Fisher procedure, i. e. state positive or negative dependence or make no statement according as

$$H(X) \leq \varepsilon \ , \quad H(X) \geq 1 - \varepsilon \ , \quad \varepsilon < H(X) < 1 - \varepsilon \ , \tag{10}$$

where

$$H(x) = \sum_{y=0}^{x} h(y) \ ; \quad h(x) = \binom{M}{x} \binom{n-M}{L-x} / \binom{n}{L} \tag{11}$$

(More precisely the optimum, almost unique solution is the following. Let c_1 and c_2 be integers and $0 \leq \gamma_1 < 1$, $0 \leq \gamma_2 < 1$ be such that

$$\begin{aligned} H(c_1 - 1) + \gamma_1 h(c_1) &= \varepsilon \\ 1 - H(c_2) + \gamma_2 h(c_2) &= \varepsilon \end{aligned} \tag{12}$$

Thus $c_1, c_2, \gamma_1, \gamma_2$ depend on the marginals in (9) , M, L, n . Then state negative dependence if $X < c_1$ or if $X = c_1$ and an event with probability γ_1 occurs. State positive dependence if $X > c_2$ or $X = c_2$ and an event with probability γ_2 occurs. Otherwise make no statement. Of course, the randomization will never be used in practice).

For large n the Irwin - Fisher test is approximately equal to the chi-square-goodness of fit test, i. e. examine first if

$$\chi^2 = \frac{(LM-nX)^2}{nL(n-L)} \geq z_{1-2\varepsilon} \tag{13}$$

where $z_{1-2\varepsilon}$ is the $1-2\varepsilon$ fractile of the chi-square distribution with 1 degrees of freedom. If such is the case then state negative or positive dependence according as

$$X < \frac{M}{n}L \quad \text{or} \quad > \frac{M}{n}L \tag{14}$$

Now, the last part of the procedure is trivial since

$\frac{M}{n}L = E(X|L,M)$ is the expected number of $A_1 \cap B_1$ (conditionally) under independence. Thus the "significance testing" (13), i.e. the "testing of the null hypothesis of independence" is the important part of the procedure. Again there is nothing contradictory in

making such a test even if it is a priori known that independence is precluded. It is numerical convenience only which dictates such a procedure. The statistical problem and the actual procedure can be formulated without any reference to a "null-hypothesis".

It is noteworthy that if the construction of the optimum test had been performed in two steps, the first being to construct a uniformly most powerful unbiased test of $\xi = 0$; then a different test would arise, as shown by Erling Sverdrup [4].

3 Testing several parameters.

a. Multiple comparison in linear-normal trials.

As a prototype of the general situation to which the theory applies let us consider the analysis of variance in the one-way lay-out. X_{ij}; $i = 1,2,\ldots, n_j$; $j = 1,2,\ldots p$, are independently and normally distributed with $\mathrm{var}X_{ij} = \sigma^2$ (unknown) and $Ex_{ij} = \xi_j$ (unknown). We are interested in comparing the different ξ_j; e. g. in pairs $\xi_i - \xi_j$, or if one group of ξ_j on the average in greater than another group, or if ξ_j is covariant with some quantity t_j ($\Sigma\xi_j(t_j - \bar{t})$), or if the influence of t_j on ξ_j is accelerating

$(\xi_{i+1}-\xi_i)/(t_{i+1}-t_i) - (\xi_i-\xi_{i-1})/(t_i-t_{i-1}) > 0$, etc. etc.

In short, we are interested in discovering contrasts $\sum_{j=1}^{p} c_j\xi_j$ ($\Sigma c_j = 0$) which are > 0. According to Scheffé's well-known method it should be asserted that $\sum_{j=1}^{p} c_j\xi_j > 0$ if

$$\Sigma c_j\bar{X}_j > \sqrt{(p-1)f}\,S\,\sqrt{\Sigma c_j^2/n_j} \qquad (15)$$

where $\bar{X}_j$ is the class average, S^2 is the usual unbiased estimate of σ^2 with $n-p$ degrees of freedom and f is the $1-\varepsilon$ fractile of the Fisher distribution with $p-1$ and $n-p$ degrees of freedom ($n = \Sigma n_j$).

Obviously an error is committed if there exists a $c = (c_1, \ldots, c_p)$ such that (15) is true whereas $\Sigma c_i\xi_j \leq 0$.

Now, the fundamental property of Scheffé's method is that the probability of committing at least one such error is at most ε for any value of $(\xi_1,\ldots, \xi_p)$.

Note that in this formulation of the multiple comparison method and its property no nullhypothesis need be mentioned.

There are three reasons why it may be convenient to bring in the conventional nullhypothesis $H_0 : \xi_1 = \ldots = \xi_p$. The first reason is that contrasts really say something about how $\xi_1, \ldots, \xi_p$ deviate from H_0. The second reason is that the above mentioned probability of committing at least one error assumes its maximum for $\xi_1 = \ldots = \xi_p$.

The third reason is the following. It is obviously inconvenient to look for a $(c_1, \ldots, c_p)$ for which (15) is true. It may be very few of them, perhaps none, in cases where the observations contain little information. Hence it is convenient that we have the following purely algebraical relationship. (15) is true for some $(c_1, \ldots, c_p)$ if and only if

$$F = \Sigma n_j (\bar{X}_j - \bar{X})^2 / (p-1)S^2 > f \qquad (16)$$

(where $\bar{X}$ is the total mean). If, therfore, $F \leq f$ it is futile to look for significant contrasts. Thus, as a computational rationalization, the Fisher F -test (16) should be performed first, i. e. one should test the nullhypothesis $\xi_1 = \ldots = \xi_p$ regardless of whether one has any belief in this hypothesis or not, and even if it is known that this hypothesis is wrong. Thus again, the significance test is a kind of clearing test in a multiple inference problem, significance signifying that inferences can be drawn.

There are other useful methods for multiple comparison beside the Fisher-Scheffé-test treated above. In the case where $n_1 = n_2 = \ldots = n_p = m$, the Student-Tukey's method is to the effect that it should be asserted that $\Sigma c_i \xi_i > 0$ whenever

$$\Sigma c_i \bar{X}_i > vS\Sigma|c_i|/2\sqrt{m} \qquad (17)$$

where v is $1 - \varepsilon$ fractile for the Studentized range with p variables and $p(m-1)$ degrees of freedom. Then the probability of an error is at most ε for any $\xi_1, \ldots, \xi_p, \sigma$. The corresponding "significance test" is in this case

$$\max \bar{X}_i - \min \bar{X}_i > vS/\sqrt{m} \qquad (18)$$

Relatively to the $\binom{p}{2}$ paired comparisons of ξ_j-s, i. e. to the contrasts $\xi_i - \xi_j$, the test has the following optimum property (which is an easy consequence of Lehmann [3]).

For any decision d and any $\xi = (\xi_1,\ldots,\xi_p)$ let

$$F(\xi,d) = \text{number of false statements}$$
$$U(\xi,d) = \text{number of neglected contrasts } \xi_i - \xi_j > 0 \qquad (19)$$
$$R(\xi,d) = \text{number of true statements}$$
$$F + U + R = \binom{p}{2}$$

If e.g. $\xi_1 > \xi_2 > \xi_3 > \xi_4 > \xi_5$ $(p = 5)$ and the statement d is "$\xi_2 > \xi_1,\xi_3,\xi_4,\xi_5$ and $\xi_3 > \xi_4$ and $\xi_1 > \xi_4$", then $F = 1$, $U = 4$, $R = 5$. We now define the loss of making decision d as

$$L(\xi,d) = (a+b)F(\xi,d) + bU(\xi,x) \qquad (20)$$

where $a \geq b > 0$. Then with $\epsilon = b/(a+b)$ the Student-Tukey-test is the uniformly least risky among all risk-unbiased tests.

It is of some interest to point out that the comparison of the population variances in the one-way-lay-out (assuming that they can be different) could be carried out in a manner quite similar to the comparison of means by the Student-Tukey's method; i.e. if the p estimates of the variances are $S_1^2,\ldots,S_p^2$, then the variance in group i is declared greater than the variance in group j if $S_i^2/S_j^2 >$ some constant. This is Hartley's method. It is also optimum in the sense explained above relatively to the loss function (19) with $\xi_1,\ldots,\xi_p$ replaced by $\sigma_1^2,\ldots,\sigma_p^2$. The corresponding significance test is $\max S_i^2/\min S_i^2 >$ some constant. This is different from the Bartlett's test based on the likelihood ratio principle. Bartlett's test is a result of the common tendency among statisticians to press any situation into a two-decision problem. The excuse for doing so has been that it simplifies the matter. That this is not always the case is Bartlett's test an example of. It is relatively complicated and of doubtful practical value. Hartley's test is much simpler and is based on a more reasonable way of posing the problem. It is rather peculiar that the statisticians have often had qualms of conscience when applying such methods.

Thus H.O. Hartley [6] calls his method a "short cut" test and I.W. Tukey [7] talks about "quick and dirty methods". The practical intuition of the statistician leads him to feel that such methods are to be preferred and he attempts to justify this preference with the amount of computational work, despite the fact that this could obviously not be the motive. The real reason has been given above.

b. Chi-square goodness of fit test in multinomial trials.

After what has been said about double dichotomies and Fisher-Scheffé's test, our attitude to the classical Karl Pearson chi-square goodness of fit test is obvious. In each of n independent trials one of r exclusive events $A_1, A_2, \ldots, A_r$ occurs with probabilities $p_1, p_2, \ldots, p_r$, respectively $(\Sigma p_j = 1)$.

We want to decide in which manner $p_1, \ldots, p_r$ deviate from

$$p_i = p_i(\theta) \; ; \quad i = 1,2,\ldots,r \tag{21}$$

where $p_i(\theta)$ are some specified functions of an unknown parameter $\theta = (\theta_1, \ldots, \theta_s)$. To be more precise, given that there exists no θ such that (21) is true, give a description of interesting deviations between the true p_i and the form of $p_i(\theta)$ for any θ.

The classical procedure is the following. Find first an estimate $\hat{\theta}$ of θ (e.g. by the maximum likelihood method), then verify if

$$\chi^2 = \sum_i (X_i - np_i(\hat{\theta}))^2 / np_i(\hat{\theta}) > z \; , \tag{22}$$

where $X_1, \ldots, X_r$ are the number of times $A_1, \ldots, A_r$ respectively occur in the trials, and z is the $(1-\varepsilon)$-fractile of the chi-square-distribution with $r-s-1$ degrees of freedom. If (22) holds then

compare $X_1,\ldots, X_r$ with $np_1(\theta^2),\ldots, np_r(\theta^2)$ and place confidence in the most conspicuous interesting deviations. By this method the probability of making an error is at most ε .

The double dichotomy example above is an instance in point. Another instance is the distribution of births over 12 months of the year; $X_1,\ldots, X_{12}$ being the number of births respectively, $\sum_1^{12} X_j = n$. If, after looking at the data, X_2 is found to be very high, this should of course not be tested by comparing X_2/n to $1/12$; but by testing if

$$\sum_1^{12} \frac{r}{n}(X_i - \frac{n}{r})^2 \qquad (23)$$

is large with 11 degrees of freedom.

Of course it is unsatisfactory that at present no precise method exists for performing the multiple decision part of the procedure, but that should not deter professional statisticians from recommending the method. The remarks about the "null-hypothesis" (21) are similar to those made in connection with the previous situations treated here. The hypothesis is anchorage point, relatively to which interesting statements could be made.

4. Curve-fitting by means of "wrong" analytical expressions.

Perhaps it would be appropriate in the above context to say something about this old statistical problem, much discussed among actuaries.

We shall illustrate the problem by the example of analysis variance in section 2b above.

Assume that we have discovered, by looking at the group means $\bar{X}_i$; $i = 1,2,\ldots,p$, that ξ_i varies roughly linearly with some quantity t_i, $i = 1,2,\ldots, p$. Then we might be interested in joint confidence intervals for all

$$\eta(t) = \alpha + \beta(t - \bar{t}) \qquad (24)$$

when t varies continuously. Here

$$\bar{t} = \frac{1}{n}\Sigma n_j t_j \ , \quad \alpha = \frac{1}{n}\Sigma n_j \xi_j \ , \quad \beta = \Sigma n_j(t_j - \bar{t})\xi_j / \Sigma n_j(t_j - \bar{t})^2 \qquad (25)$$

This is the famous problem of Working and Hotelling [5], 1929.

Let

$$M = \Sigma n_j(t - \bar{t})^2 \ , \quad a = \frac{1}{n}\Sigma n_j \bar{X}_j \ , \quad b = \Sigma n_j(t_j - \bar{t})^2 \bar{X}_j / M \ ,$$

$$\hat{\eta}(t) = a + b(t - \bar{t}) \ , \ c_j(t) = \sqrt{n_j}\left[\frac{1}{n} + \frac{t_j - \bar{t}}{M}(t - \bar{t})\right] \qquad (26)$$

$$K_t^2 = pf \cdot \sum_{j=1}^{p} c_j^2(t)$$

where f is the $(1 - \varepsilon)$ - fractile of the Fisher distribution with p (not $p - 1$) and $n - p$ degrees of freedom. Then by using Scheffé's multiple comparison method we find that

$$\hat{\eta}(t) - K_t S < \eta(t) < \hat{\eta}(t) + K_t S \qquad (27)$$

defines a $(1-\varepsilon)$-confidence band for the regression values. (If we were interested in a particular $\eta(t)$, we would of course have used t_o^2 instead of pf, where t_o is the $(1 - \frac{\varepsilon}{2})$ - fractile of the Student distribution with $n - p$ degrees of freedom).

Note that we do <u>not</u> assume $EX_{ij} = \alpha + \beta(t_j - \bar{t})$. The statement above that "ξ_j varies roughly linearly with t_j" is just a motivation, it is not a basis for the mathematical derivation of the method. If we had assumed that $\xi_j = \alpha + \beta(t_j - \bar{t})$, then we could get a confidence band.

$\eta(t) \in \hat{\eta}(t) \pm \sqrt{2f(\frac{1}{n} + (t-E)^2 S/M}$, where f is now the $(1 - \varepsilon)$ - fractile of the Fisher distribution with 2 and $n - 2$ degrees of freedom. This was Hotelling's and Working's solution.

References

[1] Erich Lehmann: "Significance level and Power" (1958) Ann. Math. Stat. vol 29 p. 1167.

[2] Anders Hald: "The size of the Bayes and minimax tests as function of the sample size and loss ration". Skand. Akt. tidsskrift (1971) p. 74.

[3] Erich Lehmann: "A Theory of multiple decision problems"(1957), (1957), Ann. Math. Stat. vol. 28. Part I on p. 1 and part II on p. 547.

[4] Erling Sverdrup: "Similarity, unbiasedness, minimaxibility and admissibility of statistical test procedures. Skand. Akt. tidsskrift (1953).

[5] H. Working and H. Hotelling: "Application of the theory of error to the interpretation of trends". J. Am. Statist. Assoc., Suppl. (Proc) vol.24, p. 73 (1929).

[6] H.O. Hartley: "The maximum F-ratio as a short-cut test for heterogenity of variances". Biometrika, vol. 37 (1950), p. 308.

[7] I.W. Tukey: "Quick-and-dirty methods in statistics, part II, Simple analysis of standard designs". Am. Soc. Qual. Cont., 5th Ann. Conv. Trans.(1951) p. 189.